Шарова Т. В., соискатель
Московских В. А., доц., канд. экон. наук
Гольдштейн С. Л., проф., д-р техн. наук

ТЕХНОЛОГИЯ АНАЛИЗА ТЕКСТОВ: ТИПОЛОГИЯ ПОНЯТИЙ

Актуальность и постановка задачи

Развитие систем, основанных на знаниях, требует переосмысления всех предыдущих достижений в области работы с инфосырьем. Одной из опорных точек при этом могут быть технологии анализа текстов. Состояние, развитие, успехи и проблемы в этой области имеют многочисленную библиографию [1-42]. Список ключевых слов при этом включает в себя термины: текст [1,2], текстология [3, 6], теория текста [2], критика текста [5], текстовый анализ [1], лингвистика [7], семиотика [8, 9, 31], синтаксис [8, 10], семантика [8], прагматика [8, 10], герменевтика [11-14], экзегетика [15], контент-анализ [16-28], риторика [7, 9], кинесика [7] и т.д. Единого взгляда на этот конгломерат понятий нет, нет и типологии. Сформировалось несколько отдельных теоретических аспектов: философский, литературный, филологический и т.п. Привлекаются разные парадигмы: гуманитарная, техническая, естественно-научная. Практические аспекты отличаются объектами приложения (художественная литература, религиозные тексты, научные труды и т.д.), субъектами (автор текста, редактор, издатель, читатель) и методиками анализа (от вербально-описательных до строгих математических). Актуальность наведения хотя бы субпорядка в этой области очевидна.

Тезаурус основных понятий как база информационного субпорядка

По ключевым словам просмотрено 50 библиографических источников за период с 1950 по 2006 гг. и 45 адресов сети Internet, оценены полнота и достоверность информации. В результате ее анализа построены фрагменты иерархических тезаурусов по нескольким основным понятиям (рис. 1 - 11).

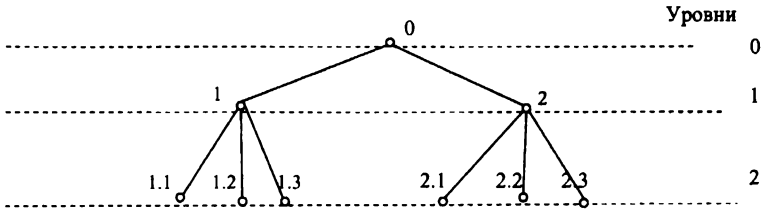

Рис. 1. Фрагмент иерархии понятий по термину «Коммуникация»
(0 – коммуникация, 1 аспекты, 2 – типы, 1.1 – технический,
1.2 – семантический, 1.3 – прагматический, 2.1 – человек / человек,
2.2 – человек / ЭВМ, 2.3 – ЭВМ / ЭВМ)

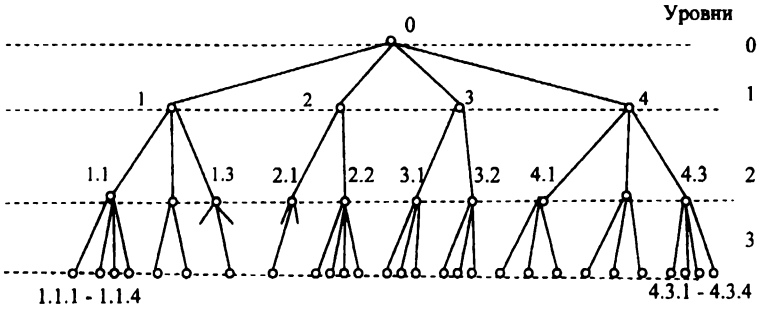

Рис. 2. Фрагмент тезауруса понятий по термину «Теория текста»

(0 – теория текста: 1 – лингвистика; 1.1 – внешняя лингвистика; 1.1.1 – язык в связи с историей народа и цивилизации; 1.1.2 – язык в связи с политикой; 1.1.3 – язык в связи с литературой; 1.1.4 – язык в связи с его географическим распространением и т.д.; 1.2 – внутренняя лингвистика (языкознание); 1.2.1 – устройство языка; 1.2.2 – структура языка; 1.3 – интерлингвистика; 1.3.1 – международный язык как средство межъязыкового исследования; 2 – структура текста; 2.1 – графическое деление текста; 2.1.1 – рубрикация; 2.2 – смысловая организация текста; 2.2.1 – информационная; 2.2.2 – логическая; 2.2.3 – психологическая; 2.2.4 – эстетическая; 3 – герменевтика (толкование); 3.1 – традиционная классическая герменевтика; 3.1.1 – синтаксический анализ; 3.1.2 – семантический анализ; 3.1.3 – прагматический анализ; 3.2 – современная литературная герменевтика 3.2.1 – анализ внутренней логики единой конструкции текста; 3.2.2 – анализ значения текста; 3.2.3 – контент–анализ текста; 4 – грамматика; 4.1 – словообразование; 4.1.1 – слово как отдельная единица; 4.1.2 – словообразовательная система; 4.1.3 – способы словообразования; 4.2 – морфология; 4.2.1 – морфология слова; 4.2.2 – морфология грамматических изменений слова; 4.2.3 – морфология грамматических характеристик слова; 4.3 – синтаксис; 4.3.1 – синтагматика слова; 4.3.2 – синтаксис словосочетания; 4.3.3 – синтагматика предложения; 4.3.4 – синтаксис форм слова)

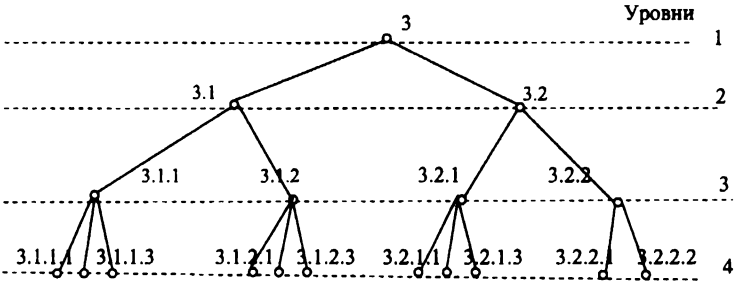

Рис. 3. Вариант тезауруса понятий по термину «Предметно-инструментальная база герменевтики

(3.1 – предмет, 3.2 – инструмент, 3.1.1 – содержательный аспект, 3.1.2 – временной аспект, 3.2.1 – метод, 3.2.2 – сознание, 3.1.1.1 – литературный, 3.1.1.2 – филологический, 3.1.1.3 – философский, 3.1.2.1 – ренессанс, 3.1.2.2 – классика, 3.1.2.3 – современность, 3.2.1.1 – дескриптивный, 3.2.1.2 – нормативный, 3.2.1.3 – исторический, 3.2.2.1 – толкователя, 3.2.2.2 – читателя)

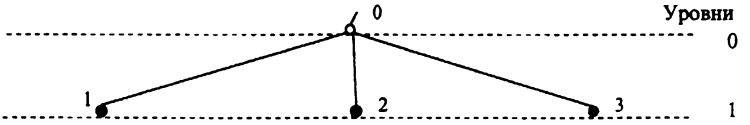

Рис. 4. Начало тезауруса понятий по термину «Контентный анализ»

(1 – типология текстов для анализа, 2 – задачи анализа, 3 – методы анализа)

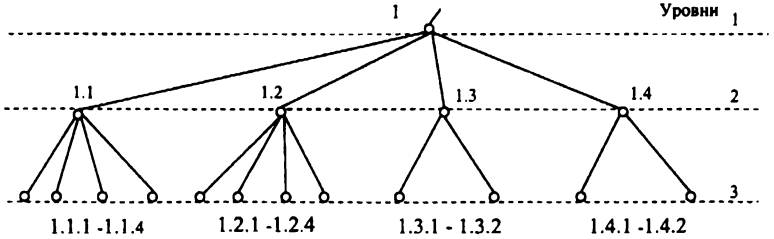

Рис. 5. Тезаурус понятий по классификации текстов в составе контентного анализа

(1 – типология текстов для анализа: 1.1 – по жанру: 1.1.1 – научные, 1.1.2 – художественные 1.1.3 – общественно-политические; 1.1.4 – информационно-документальные; 1.2 – по способу фиксации информации: 1.2.1 – письменные; 1.2.2 – устные; 1.2.3 – фонетические (рассчитанные на

слуховое восприятие); 1.2.4 – аудиовизуальные; 1.3 – по целевому назначению: 1.3.1 – естественно функционирующие (цели внешние по отношению к исследованию); 1.3.2 – «целевые» документы (полученные согласно исследовательской программе); 1.4 – по числу авторов: 1.4.1 – один автор / составитель); 1.4.2 – два и более авторов / составителей)

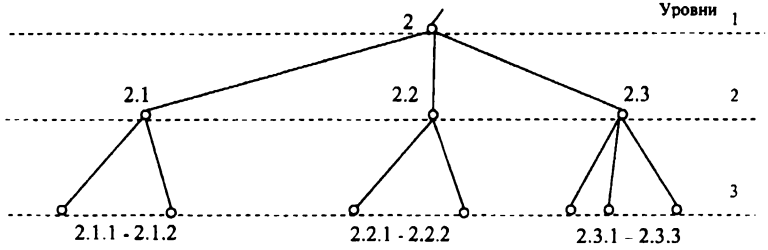

Рис. 6. Тезаурус понятий по задачам анализа текстов в составе контентного анализа

(2 – задачи анализа текстов: 2.1– задачи извлечения информации; 2.1.1 – извлечения открытой для понимания информации; 2.1.2 – извлечения суггестивной (внушенной) информации; 2.2 – задачи анализа информации по результату; 2.2.1 – реферирования; 2.2.2 – экспертизы; 2.3 – задачи анализа по предмету; 2.3.1 – интегральный анализ; 2.3.2 – ситуационный анализ; 2.3.3 – проблемный анализ).

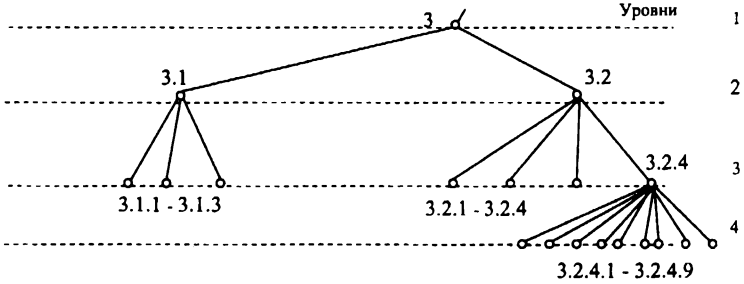

Рис. 7. Тезаурус понятий по методам анализа текстов в составе контентного анализа

(3 – методы анализа текстов; 3.1 – традиционные; 3.1.1– статистический анализ; 3.1.2 – лингвистический анализ; 3.1.3 – морфологический анализ; 3.2 – контекстные; 3.2.1 – информационный анализ; 3.2.2 – структурный (тектонический) анализ; 3.2.3 – проблемно-мотивационный анализ; 3.2.4 – контент-анализ, 3.2.4.1 – качественный контент-анализ; 3.2.4.2 – количественный контент-анализ; 3.2.4.3 – контент-анализ простых частот; 3.2.4.4 – контент-анализ относительных частот; 3.2.4.5 – контент-анализ категорий; 3.2.4.6 – норма-анализ; 3.2.4.7 – контент-анализ связей категорий; 3.2.4.8 – контент-анализ контекстный; 3.2.4.9 – автоматическая категоризация)

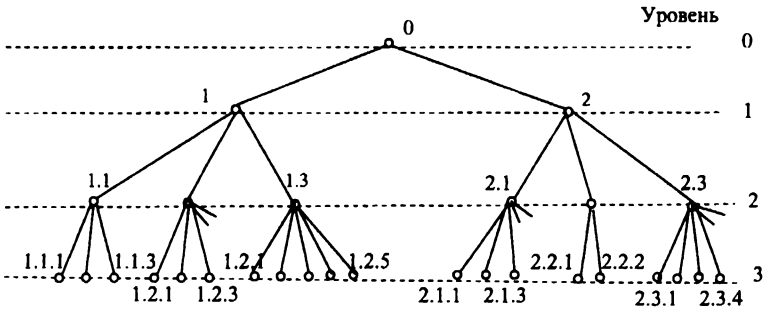

Рис. 8. Иерархическая модель риторики

(0 – риторика; 1 – античная риторика: 1.1 – классификация по источникам красноречия: 1.1.1 – дарование; 1.1.2 – обучение; 1.1.3 – упражнение; 1.2 – по целям красноречия; 1.2.1 – убедить; 1.2.2 – усладить; 1.2.3 – взволновать; 1.3 – по этапам разработки речи: 1.3.1 – нахождение материала; 1.3.2 – расположение материала; 1.3.3 – словесное выражение материала; 1.3.4 – запоминание; 1.3.5 – произнесение; 2 – стилистика; 2.1 – по видам стилистики: 2.1.1 – стилистика художественной литературы; 2.1.2 – сопоставительная стилистика; 2.1.3 – историческая стилистика; 2.2 – по видам стилистического значения: 2.2.1 – функциональное; 2.2.2 – экспрессивное; 2.3 – классификация по разделам: 2.3.1 – функциональная; 2.3.2 – языковых единиц; 2.3.3 – текста; 2.3.4 – художественной речи, …)

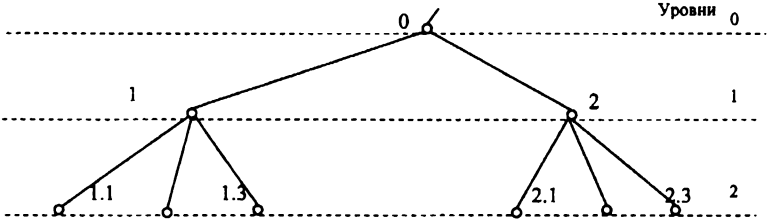

Рис. 9. Иерархическая модель семиотики

(0 – семиотика, 1.1 – классификация по разделам изучения: 1.1 – синтаксис; 1.2 – семантика; 1.3 – прагматика; 2 – классификация по основным способам изучения; 2.1 – абстрактная: 2.2 – теоретическая; 2.3 – эмпирическая)

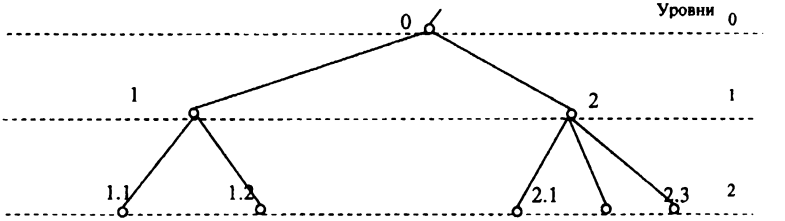

Рис. 10. Иерархическая модель кинесики

(0 – кинесика, 1 – по видам жестов: 1.1 – универсальная; 1.2 – социально-обусловленная; 2 – по основным функциям; 2.1 – вспомогательного вида общения: 2.2 – сопровождения речи; 2.3 – субститута речевых отрезков)

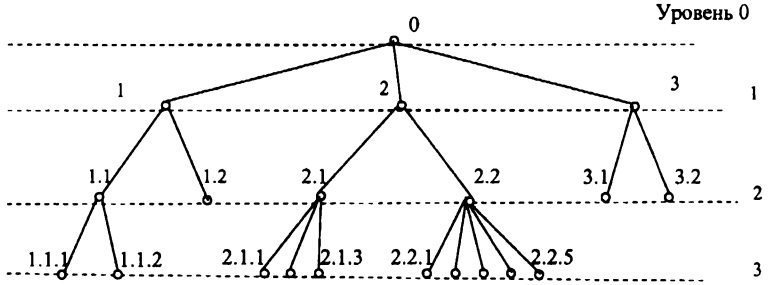

Рис. 11. Иерархическая модель понятий по термину «Математическая лингвистика»

(0 – математическая лингвистика, 1 – методы, 2 – теории, 3 – модели, 1.1 – математической логики, 1.2 – алгебры, 2.1 – синтаксических структур, 2.2 – формальных грамматик, 3.1 – аналитические, 3.2 – дешифровочные, 1.1.1 – теории алгоритмов, 1.1.2 – теории автоматов, 2.1.1 – системы составляющих, 2.1.2 – дерева синтаксического подчинения, 2.1.3 – системы синтаксических групп, 2.2.1 – порождающей грамматики Хомского, 2.2.2 – доминационной грамматики, 2.2.3 – грамматики синтаксических групп, 2.2.4 – грамматики деревьев, 2.2.5 – грамматики Монтегю)

Предлагаемый тезаурус

Предлагаемый тезаурус приведен на рис. 12.

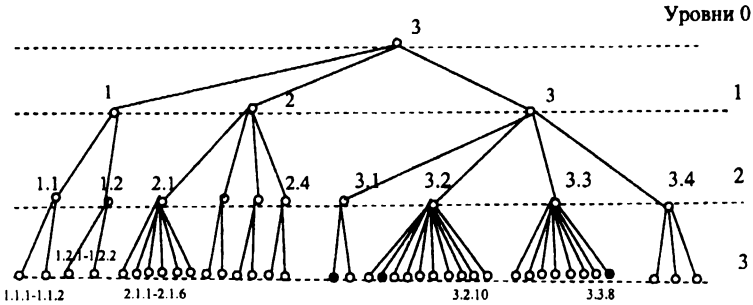

Рис. 12. Предлагаемая иерархия понятий по термину «Коммуникации»

(1 – коммуниканты, 2 – тексты, 3 – диалог в парадигмах, 1.1 – неживой природы, 1.2 – живой природы, 2.1 – по жанру, 2.2 – по способу фиксации информации, 2.3 – по целевому назначению, 2.4 – по авторам: 3.1 – системной, 3.2 – естественно-научной, 3.3 – гуманитарной, 3.4 – бытовой; 1.1.1 – первой природы, 1.1.2 – второй природы, 1.2.1 – с первой сигнальной системой, 1.2.2 – со второй сигнальной системой (речью), 2.1.1 – научные, 2.1.2 – художественные, 2.1.3 – технические, 2.1.4 – общественно-политические, 2.1.5 – документально-информационные, 2.1.6 – бытовые, 2.2.1 – письменные, 2.2.2 – устные, 2.3.1 – по типу, 2.3.2 – по виду, 2.4.1 – рангу, 2.4.2 – количеству; 3.1.1 – в системологии, 3.1.2 – в системотехнике, 3.2.1 – в математике, 3.2.2 – в информатике, 3.2.3 – в физике, 3.2.4 – в химии, 3.2.5 – в астрономии, 3.2.6 – в науках о земле, 3.2.7 – в биологии, 3.2.8 – в экологии, 3.2.9 – в экономике, 3.2.10 – в медицине, 3.3.1 – в философии, 3.3.2 – в психологии, 3.3.3 – в социологии, 3.3.4 – в истории, 3.3.5 – в культурологии, 3.3.6 – в политологии, 3.3.7 – в юриспруденции, 3.3.8 – в теории текста, 3.4.1 – дом, 3.4.2 – одежда, 3.4.3 – пища…)

Видно, что рис. 12 может быть взят за базу типологии понятий по технологии анализа текстов. При этом особый интерес представляют диалоги (вершины 3.1.1, 3.2.2, 3.3.8).